Assessing Researcher Interdisciplinarity: A Case Study of the University of Hawaii NASA Astrobiology Institute

Michael Gowanlock & Rich Gazan

M. Gowanlock
Department of Information & Computer Sciences and University of Hawaii NASA Astrobiology Institute, University of Hawaii
POST 310, 1680 East-West Road, Honolulu, HI 96822, USA
gowanloc@hawaii.edu

R. Gazan
Department of Information & Computer Sciences, Library & Information Science Program and University of Hawaii NASA Astrobiology Institute, University of Hawaii, Hamilton Library 2H, 2550 McCarthy Mall, Honolulu, HI 96822, USA
gazan@hawaii.edu

Abstract

In this study, we combine bibliometric techniques with a machine learning algorithm, the sequential Information Bottleneck, to assess the interdisciplinarity of research produced by the University of Hawaii NASA Astrobiology Institute (UHNAI). In particular, we cluster abstract data to evaluate ISI Web of Knowledge subject categories as descriptive labels for astrobiology documents, and to assess individual researcher interdisciplinarity to determine where collaboration opportunities might occur. We find that the majority of the UHNAI team is engaged in interdisciplinary research, and suggest that our method could be applied to additional NASA Astrobiology Institute teams to identify and facilitate collaboration opportunities.

Keywords: Astrobiology, Bibliometrics, Information bottleneck method, Interdisciplinary science, Machine learning, Text mining

1 Introduction

Astrobiology, the study of the origin, evolution, distribution, and future of life in the universe, is a relatively new field comprised of researchers from a range of scientific disciplines. Apart from its sublime object of study, astrobiology has been identified as a field that can integrate diverse sciences (Staley, 2003), provide a tangible target for interdisciplinary science education (Cockell, 2002), and provide a pathway to adult science literacy (Oliver and Fergusson, 2007). Many of the field's core questions require knowledge from multiple disciplines to be harvested, integrated and applied outside their source domains, and as such, astrobiology is inherently interdisciplinary. For example, the University of Hawaii NASA Astrobiology Institute (UHNAI) studies the origin of water in the solar system and beyond, in the context of understanding the origins of life. Astronomers, chemists, geologists, oceanographers and biologists work together to study data from meteorite fragments to comets to the interstellar medium to address the question of where else in the universe water, and thus life, might be found. Without collaboration across disciplinary boundaries to interpret often-scarce data, important questions in astrobiology will remain incompletely addressed. Developing a method to identify, measure and catalyze interdisciplinary work in the astrobiology research environment is the goal of this paper.

One of the benefits of a broad-based research community is that new developments in astrobiol-

ogy occur fairly frequently. The downside is that researchers must stay abreast of these numerous developments both inside and outside of their home fields. As new astrobiology research findings are reported, the considerable effort involved in finding, evaluating and integrating them indicates a need for a better understanding of how findings in one field might inform others, and to identify potential collaboration opportunities between individual researchers working on similar questions.

Our previous example suggested that knowledge from multiple disciplines is required to understand the origin of water to answer questions regarding the origin of life. Satisfactorily understanding the research record of scientists that work in this area requires measuring interdisciplinarity on an acute scale. Therefore, instead of exploring large-scale trends in publications using a top-down approach, it is necessary to have a large amount of data that represents the research track of each astrobiologist using a bottom-up approach.

A common method used to examine the potential of collaboration across disciplinary boundaries is to interview domain experts, but this method suffers from several limitations, such as sample size and subjectivity problems (Zhang et al., 2011). Furthermore, given that the subject matter of astrobiology spans many disciplines, meaningful analysis of the responses would require the knowledge of an astrobiology polymath. After considering these limitations, we suggest that measuring interdisciplinarity should be guided by one or more individuals versed in astrobiology, but whose expertise need not span all of its constituent disciplines. Therefore, an unsupervised approach is optimal as such methods can find trends in data without prior knowledge of its structure.

As of 2011, the NASA Astrobiology Institute is comprised of 14 teams spanning ten universities in addition to NASA Ames, Goddard, and the Jet Propulsion Laboratory. While a cross-team analysis is beyond the scope of this paper, we suggest that our method for measuring researcher interdisciplinarity at UHNAI could be extended to other NAI teams, and to scientific collaborations more broadly. Furthermore, our method suggests where collaborations might productively occur, and allows us to better understand the nature of interdisciplinary scientific discovery.

In this pilot study, we investigate the use of an unsupervised machine learning clustering techique, the sequential Information Bottleneck (sIB) (Slonim et al., 2002) to aid in measuring researcher interdisciplinarity. Furthermore, we assess whether Journal Subject Categories from the Thomson Reuters Web of Knowledge database suite are sufficient for labelling astrobiology documents. The clustering and classification of text allow interdisciplinary analysis that 1) describes

collaboration and the integration of knowledge and 2) draws conclusions that are useful to astrobiology researchers by uncovering the underlying structure of research tracks. The results of this pilot study will serve to guide a subsequent investigation that will identify collaboration opportunities and measure the disciplinary roots across the entire NASA Astrobiology Institute.

Researchers in astrobiology tend to be comfortable speaking in the language of multiple scientific disciplines. As suggested in Gargaud and Tirard (2011), these interdisciplinary researchers are somewhat isolated from their counterparts in other academic departments. The multidisciplinary context given by astrobiology affords an excellent opportunity to examine the methods used to study researcher interdisciplinarity and knowledge integration. Furthermore, we aspire to use this information in an iterative process to facilitate collborations between researchers, and ease the cognitive load of a single researcher who wishes to integrate knowledge from multiple disciplines.

2 Background

Research that occurs at the intersection between disciplines is thought to lead to great advances in science (Porter and Rafols, 2009). Many funding agencies exist specifically to support and encourage interdisciplinary research; the U.S. National Science Foundation's interdisciplinary research efforts span all of their divisions and directorates (National Science Foundation, Accessed November 21, 2011). For example, some authors measuring interdisciplinarity lament that there is not enough coverage of the societal causes for climate change (Bjurström and Polk, 2011) as described in the Intergovernmental Panel on Climate Change (IPCC) literature. In this specific case, measuring both the disciplinary diversity and the integration of knowledge is of paramount importance to ensure that future IPCC reports include appropriate factors. A cynical disposition to this problem is eloquently stated in Brewer (1999): "The world has problems, but universities have departments."

Many important terms in this work have so far been discussed without qualification. The term interdiscipinary tends to be tacitly understood by researchers, with no consensus definition. We adopt the definition suggested by Porter et al. (2007), which followed the definition given by the National Academies (2005): interdisciplinary research requires an integration of concepts, theories, techniques and/or data from two or more bodies of specialized knowledge. Multidisciplinary research may incorporate elements of other bodies of specialized knowledge, but without interdis-

ciplinary synthesis (Wagner et al., 2011) that leads to research that is greater than the sum of its parts.

Despite the increase in claimed interdisciplinarity, traditional indicators are of questionable value in assessing and quantifying interdisciplinary research (Morillo et al., 2001). Additionally, policies regarding interdisciplinarity are often based more on conventional wisdom than empirical studies (Rafols and Meyer, 2010). The usefulness of bibliometric indicators depends critically on the level at which we wish to understand the integrative process. For example, funding agencies may only require high-level publication co-authorship and collaboration statistics, describing the research performed by their grantees and the diversity of their home disciplines, but not addressing the essential aspect of synthesis. When there is no mechanism to identify, measure and encourage these points of intellectual crossover, there is no way to quantify the extent to which interdisciplinary science is taking place.

Top-down approaches have been used to map scientific literature (for example, see Boyack et al. (2005)), and often represent broad areas of science with Web of Knowledge (WoK) subject categories (SCs). For example, van Raan and van Leeuwen (2002) and Porter et al. (2007) used SCs in their methodology to measure interdisciplinarity. In these studies, SCs have been employed as de facto disciplinary boundaries, and as a benchmark to measure how much a given author, journal or research area crosses scientific fields. Unfortunately, low-level conclusions that might inform potentially productive individual collaborations cannot be made when relying on these top-down approaches, as they focus on past outputs rather than future integration. Conversely, bottom-up bibliometric approaches incorporate the authors' own words, in free-text fields such as: titles, abstracts, keywords[1] and the full text of a document. Clustering bibliometric data at this level can describe the structure of a researcher, journal or an entire field, and suggest productive future directions. Kostoff (1998) describes how citation analysis can serve as a "radioactive trace" of research impacts. One limitation of cluster analysis is that "...precise disciplinary divisions are not obtained, rendering inter-cluster links misleading" (Small, 2010), but Upham and Small (2010) propose a methodology to identify emerging "research fronts", highly cited micro-specialty areas that transcend existing fields. Their method requires that the researcher not presuppose the existence of any research field, but to rely instead on a comprehensive monitoring of citations to

[1]Keywords are not always a free-text field.

identify points across disciplines where research interests intersect, echoing one goal of the present study. Both top-down and bottom-up approaches are useful in different applications. A study by Rafols and Meyer (2010) combines bottom-up and top-down approches to measure both disciplinary diversity and knowledge integration.

Measuring scientific output in bibliometric terms requires some degree of integration and normalization of the publication records of researchers, which are published in diverse formats, venues and scholarly traditions. The publication record generally includes data such as departmental affiliations, keywords, year of publication, journal, cited references, and the abstract and/or full text of the paper. This data can be compared using various bibliometric techniques to assess interdisciplinary research. While bibliometric studies tend to rely on a citation analysis, such an analysis may not be appropriate for every discipline or field. For example, a given field may tend to reference conference proceedings, websites, newspapers, or colloquia which are not as conducive to a co-citation analysis as journal articles. Due to this observation, Sugimoto (2011) suggests that studying interdisciplinarity should include publications beyond journal articles. While we agree with this position, it happens that journals are the preferred method of communication within the great majority of the fields that compose the UHNAI team; therefore, the present study is not hindered by this limitation.

3 Methodology

In this section, we outline our method for measuring interdisciplinary research. In the previous section we noted that particular bibliometric indicators are conducive to understanding research at varying levels. One of the goals of this research is to uncover the underlying structure within an astrobiology research team that undertakes interdisciplinary projects at the macro scale, but may differ in the extent of interdisciplinary work at the micro scale. To do so, we examine the text of research publications, in particular the paper abstracts. We employ a method from the field of information theory, the sIB method, to cluster our high dimensional abstract data.

An advantage of using WoK for bibliometric studies is that it provides a mapping of SCs to each journal. Given the incommensurability of other bibliometric data (for example, journals do not agree upon a common set of keywords), SCs provide a way to compare publications on the

journal level. In Zhang et al. (2010), the authors used cross-citation analysis to create seven high-level clusters of related SCs, though their analysis was somewhat confounded by the "idiosyncrasy" that a journal may be assigned to multiple SCs in WoK. In Porter et al. (2007), the authors examine the references in sets of journal articles gathered from WoK, and relate the journals to their corresponding SCs. In this approach, a more diverse set of SCs that represent a paper derived from its references indicates a higher degree of interdisciplinarity than a set of similar SCs that represent a paper.

Using the references of a paper is a reasonable approach to measure researcher interdisciplinarity. Analoguous to Porter et al. (2007), we use the references in each UHNAI publication. In particular, we combine all of the abstracts of all of the references cited by a UHNAI publication, and use these aggregated abstracts to represent each publication. These aggregated abstracts are then used to create the feature vectors supplied to the sIB clustering algorithm. Another text mining study, Kostoff et al. (2001), employed free-text fields (such as title, keywords and abstracts) of cited/citing publications in combination with phrase frequency analysis and phrase clustering analysis to obtain a low-level understanding of research impact and interdisciplinary research.

In the present study, we focus on the abstracts of cited papers only, and we do not consider the papers that cite the UHNAI papers. A major limitation of studying the citations to the UHNAI papers is that it would require the database to contain those papers that cite a particular work, which varies between disciplines/fields/databases. The same is true of those references that are cited in our UHNAI papers. To obviate this problem, we elect to use the NASA Astrophysics Data System (ADS) to collect the majority of our abstracts, as a large fraction of the UHNAI team are covered in the database. The extensive coverage in ADS ensures that a considerable majority of papers referenced by the UHNAI team are within the database. However, previous research has illustrated how the differences in scientific publication patterns between fields often require that records from multiple databases be harvested to encompass the output of interdisciplinary scientific researchers (see, for example, Kousha and Thelwall, 2008). For UHNAI authors whose publications were not sufficiently represented in ADS, we used WoK to obtain their publication data and cited references. As it turns out, those authors, and the papers they cite, were highly represented in WoK. We were able to gauge author coverage in ADS and WoK by consulting the CV of each UHNAI team member.

When working with WoK data, it is important to be mindful of the differences in institutional subscriptions, which include access to different subsets and date ranges of WoK's constituent databases, and may affect the results of bibliometric analysis (Derrick et al., 2010; Jacsó, 2005). Therefore, we provide a list of the UH WoK subscriptions at the time of data collection:

- Web of Science, 1980 -
- Biological Abstracts, 1969 -
- Medline, 1950 -
- Journal Citation Reports Science & Social Science editions, 2004 -

Before clustering the UHNAI papers and their associated abstracts, we examine whether WoK SCs are sufficient for labelling astrobiology documents. While we believe SCs are useful in mapping scientific research on a large scale, whether they are appropriate for classifying individual publications remains an open question[2]. We cluster a corpus of astrobiology abstracts labelled with their corresponding conflated SCs (described below), and assume that if a given cluster is comprised mostly of a single SC, then SCs are a sufficiently accurate classifier.

3.1 Data Collection

We gather publications by the UHNAI team members from 2001 until June 2011. Publications earlier than 2001 were not collected, as many researchers may not have been engaged in astrobiology research, and the UHNAI team was not yet founded[3]. However, we place no age restrictions on the papers that they cite.

3.1.1 NASA Astrophysics Data System (ADS)

The ADS has extensive coverage of astronomy, astrophysics and physics journal articles, pre-prints and conference proceedings. We gather the data in a semi-automated fashion. Instead of accessing

[2]The classification of documents is a requirement for an astrobiology publication information retrieval system. Our research group is inclined to create such a system. See `http://airframe.ics.hawaii.edu/` for more information.

[3]This is also the year that the journal Astrobiology began publication. While astrobiology research was, and continues to be published in other journals, this indicates that astrobiology research may not have coalesced as a field prior to 2001.

the articles through a web browser, ADS has a perl script library[4] that can be used to access parts of the database. To gather the abstracts and journals of UHNAI papers, and the abstracts and journals of the papers they cite, we employed the following procedure:

- Ran one of the ADS perl scripts to return a list of all of the publications for each UHNAI team member. This returned a list of ADS bibcodes, which uniquely identify each record in the ADS.
- Compared these papers with each author's CV to ensure that we did not collect undesired articles. For example, we filtered out papers by authors with the same last name as members on the UHNAI team.
- Used the ADS bibcodes to create a script that goes to the URL of the webpage that lists the references in each UHNAI paper. We download the individual webpages.
- Created and ran a script to capture all of the ADS bibcodes in each downloaded html web page.
- Used this list of bibcodes to get the abstracts and journals of all of the UHNAI papers and references therein using the ADS perl scripts.

3.1.2 Web of Knowledge (WoK)

The UHNAI authors whose work is underrepresented in ADS may be those who publish and cite outside astrobiology's core scientific disciplines, and contribtue to the field's interdisciplinarity. To include the published output of these researchers, and to provide a comprehensive portrait of the entire UHNAI team, we also used WoK to gather abstracts and bibliographic data. To our knowledge there is not an API or alternative way to access WoK other than using a web browser. To gather this data, we employed the following procedure:

- Created a list of all of the papers authored by UHNAI authors that were not in or underrepresented in the ADS database.
- Manually downloaded the html pages of each record describing each referenced article.

[4]The scripts can be found here: `http://vo.ads.harvard.edu/adswww-lib/`

- Created a script that parses the html pages to harvest the abstracts and journal titles.

3.2 WoK Subject Categories and Document Classification

Having collected the abstracts and journal names of UHNAI publications and references, we create a dataset that contains the abstract text and the SC of the associated journal for each UHNAI publication and the publications they cite.

Many of the SCs of the papers in our dataset were significantly underrepresented. Furthermore, as other researchers have encountered (see, for example, Zhang et al., 2010) some journals in WoK are assigned multiple SCs, necessitating some conflation into superclusters, or "macro-disciplines" (Porter and Rafols, 2009). We modify the subject categories using the following method:

- Journals with a single WoK SC that appears 10 or more times in our dataset uses the assigned WoK SC name.
- Journals with a single WoK SC that appears less than 10 times is changed to a broader WoS category (e.g. "Biochemical Research Methods" becomes "Biochemistry & Molecular Biology").
- Journals with two or more SCs of roughly the equivalent weight are assigned a new conflated SC (e.g. "Astrophysics & Geophysics").
- Journals with two or more SCs that have a clear primary SC have "-Multidisciplinary" appended to the primary name.

The ADS system also contains non-journal publications. In these instances, we manually assigned an appropriate SC to the publication. After the above procedure was performed, we ended up with 10808 abstracts in total, spanning 53 modified SCs. We eliminated those abstracts whose SCs only contributed to $\leq 0.5\%$ of the entire dataset, leaving 10359 abstracts, integrated over 13 modified SCs. There is a large class imbalance problem, as the Astronomy & Astrophysics SC significantly dominates the dataset, as shown in Table 1.

We cluster the dataset outlined above. Furthermore, we oversample the minority classes to examine if the class imbalance problem significantly affects the resultant clusters, as many data mining techniques only perform well on uniformly distributed datasets. There are a number of

methods utilized to oversample minority classes in the field of data mining. Duplicating the feature vectors in the minority classes would result in model overfitting, where a feature vector is a normalized numerical representation of the words that describe each document/instance. To alleviate this problem, we create synthetic data that is similar to the other feature vectors within a given class/SC. We use the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to produce synthetic feature vectors. A baseline dataset is created where each SC contains the same number of features as the Astronomy & Astrophysics SC (6946). We randomly sample without replacement 25% of the features contained in the baseline dataset to produce 3 subsampled datasets. While features cannot appear more than once in an individual subsampled dataset, features may be common across the 3 subsampled datasets. We cluster the data to test if our results are similar to those produced by the non-oversampled dataset.

3.3 Text Mining Aggregated Abstracts

We create a dataset of aggregated abstracts for the purposes of representing each UHNAI publication. The dataset contains 731 publications by the UHNAI team. Each publication is represented by its own abstract and the abstract of each cited publication. We aggregate all of these abstracts to represent each UHNAI publication. Note that the ADS system, and WoK to an extent, does contain publications other than journal articles, such as book chapters, conference proceedings, disserations, among others. These non-journal publications were included in the dataset, although they only constitute a small minority of the entire dataset. A large majority of the abstracts in the dataset are the same as the ones described in Section 3.2 with the exception that those abstracts that were not assigned SCs (from non-journal sources) are included in this dataset.

3.4 Text Mining and the Sequential Information Bottleneck (sIB) Method

The sIB clustering algorithm (Slonim et al., 2002) is employed to cluster our datasets described above. We chose this clustering method over others because it has been shown to perform better than other unsupervised classifiers, such as k-means (Slonim et al., 2002). Unsupervised methods aim to find groupings in the data without prior knowledge of the data's properties. It is necessary to use an unsupervised clustering method because a canonical set of astrobiology documents with which to train a clustering technique does not exist. The sIB assigns a cluster to each feature

vector. Using this method, each document must be assigned a cluster. Since no documents can be described as a hybrid of multiple clusters, we conflated multiple SCs as mentioned previously.

Our preprocessing of the datasets included converting uppercase words to lowercase, and ignoring non-alphabetical characters. We stemmed the words using the Porter stemming algorithm (Porter, 1980). We created a stopword list to remove formatting tags, and other non-content-bearing terms. We selected ∼4000 words which had a minimum frequency of 12, integrated over the entire dataset. Most of our preprocessing was performed in WEKA (Witten and Frank, 2005), and the sIB method was also executed in this environment.

We normalize each instance in our datasets (the abstract data described in Section 3.2 and Section 3.3) by creating feature vectors. Each instance/document is described by the term frequency of each word found in the ∼4000 words distilled from their respective datasets. We normalize the sum of each feature vector to 1. In the case of the aggregated abstracts, some feature vectors will be much shorter or longer than others, as there is a large range of abstract sizes, and number of references within a given publication. If we did not normalize the term frequencies, then instances with high word or low word counts may cluster together. Such clusters would be less revealing of the content of the documents themselves.

3.5 Limitations

There are several limitations to our study. First, some of the papers were authored by multiple members of the UHNAI team. In this case, we assigned the abstract data to the first-listed author on the paper, thereby not fully characterizing the research contribution of the non-primary authors. Otherwise, having multiple labels on the same document would inadvertently oversample those documents with multiple UHNAI authors. Also, there is a minor discrepancy between the abstracts gathered in ADS and WoK; ADS contains abstracts from non-journal sources, whereas WoK does to a lesser extent for the researchers studied here. The vast majority of our data was from journal articles; therefore, we do not expect this to have a significant, if any negative impact on our study. WoK maps multiple SCs to a single journal. While we need to conflate the SCs in order to compare them to clusters, the aggregation procedure undermines the fundamental function of SCs. Therefore, the results of determining if SCs are appropriate astrobiology document labels may be compromised.

4 Results

In this section we present the results of our text mining experiments.

4.1 WoK Subject Categories as Document Labels

In Figure 1 we show the results of clustering the abstract data before oversampling as described in Section 3.2. If SCs accurately reflect shared topical content of documents assigned to them, when the abstracts are clustered we should expect each SC to be primarily assigned a single cluster. However, when abstracts are assigned one of five clusters (Figure 1-top panel), we observe that the cluster membership for most SCs is heterogeneous. This observation holds for 10, 15, and 20 clusters, where the heterogeneity is even more pronounced (Figure 1). Therefore, our initial results suggest that for astrobiology abstracts, SCs do not sufficiently reflect the diverse content of publications in this domain. Figure 1 also confirms that the Astronomy & Astrophysics SC dominates the dataset, thus we oversample the data to assess the extent to which the results mentioned above hold for a uniform distribution of SCs.

Figure 2 presents the results of the oversampled data, in five clusters. Each panel reflects a separate sample, corresponding to 25% of the oversampled dataset. The cluster results of each trial are not related to each other. For example, in successive trials, the same document may be assigned to different clusters. For the Astrophysics & Geophysics SC, each trial results in different cluster assignments, though the overall distribution of clusters is roughly equal, suggesting that there is little variability between trials (Figure 2). Unlike Figure 1, some of the SCs correspond well to a single cluster, and we begin to see areas where documents across SCs are clustered together, suggesting potential interdisciplinary connection. However, since five clusters may not be sufficient to reflect the diversity of content within astrobiology, we increase the number of clusters in subsequent trials.

Doubling the number of clusters to 10 (Figure 3), one would intuitively expect more diversity in each SC; however, increasing the number of clusters also allows for the potential of each SC to dominate a single cluster. Some SCs across all three trials remain homogeneous (for example, see Biochemistry & Molecular Biology in Figure 3), suggesting that the Biochemistry and Biotechnology-related SCs are relatively effective document labels, and have the least overlap with

astrobiology's other constituent fields. The Astronomy, Oceanography and Physics SCs demonstrated somewhat less monodisciplinary dominance at the 10 cluster level; all had roughly 20% of their abstracts assigned to other clusters. The Geochemistry & Geophysics and Environmental Sciences SCs demonstrated the most diversity apart from the pure Multidisciplinary Sciences SC, though somewhat surprisingly, the Geochemistry & Geophysics-Multidisciplinary SC exhibited less diversity than its core SC.

At the 15 and 20 cluster level, as shown in Figures 4 and 5 respectively, most SCs are found in multiple clusters. This observation suggests that SCs do not map well to successively smaller clusters. For example, at the 20 cluster level, what had been homogeneous cluster membership in the Biochemistry SCs at the 10 cluster level is split into three or more clusters, neither of which is shared across any other SC. Therefore, at these clustering levels, we operationalize a "dominant cluster" as one that either constitutes 50% or more of the abstracts alone, or one that is within 50% of the size of the most common cluster[5]. By this approximation, the results at the 10-cluster level hold: as a group, the Biochemistry and Biotechnology-related SCs have the fewest dominant clusters; the Astronomy, Oceanography and Physics group slightly more, and the Geochemistry and Geophysics SCs are again the most diverse, short of the Multidisciplinary Sciences SC (Figures 4 and 5).

In some cases, the trial processes reveal some inconsistencies in the cluster membership of SCs. For example, in the Biotechnology & Applied Microbiology-Multidisciplinary SC, one would expect to have diverse membership at the 15 cluster level (Figure 4). However, trial 1 exhibits a single dominant cluster with two lesser clusters, trial 2 has two dominant and two lesser clusters, and trial 3 corresponds to a single cluster alone. While this may be an artifact of the sampling and multiple-trials processes, we would expect and find that the two related SCs, Biochemistry & Molecular Biology and Biochemistry & Molecular Biology-Multidisciplinary are found mostly within the same clusters. This observation holds for the Geochemistry & Geophysics and Geochemistry & Geophysics-Multidisciplinary SCs. The multidisciplinary SC variants (BioChem & MBio, BioChem & MBio-M and GeoChem & GeoPhys, GeoChem & GeoPhys-M) are slightly more diverse than their associated core SC, but there is a high degree of similarity between the abstracts in these two sets

[5]For example, an SC with clusters constituting 30%, 18%, 16% and 12% of the abstracts would have three dominant clusters.

of related SCs. Therefore, we conclude that even with some observed inconsistencies, the clusterer is working reliably, and may be indicating the presence of different degrees of interdisciplinarity within each SC.

4.2 Text Mining the Aggregated Abstracts

The sIB technique was employed to cluster the abstracts of the publications by the UHNAI team and the references within these publications. Figure 6, shows the results of clustering the data into 5 clusters. The results indicate that authors from their respective home disciplines cluster together. Table 2 lists the authors and their respective home disciplines. For example, the geologists Krot, Keil, Huss, Scott, and Jogo are strongly represented in cluster 4. One exception is Taylor (geologist) who clusters with the oceanographers (Cowen and Mottl). Additionally, Schörghofer (an astronomer by departmental affiliation) also clusters with the oceanographers. Furthermore, the astrochemists (Bennett and Kaiser) have all of their publications in cluster 1. This result suggests that the sIB technique is able to cluster similar research on a high-level; however, utilizing more clusters should provide a lower-level view of overlap in research interests between the authors.

When running the sIB technique for 10 clusters, we begin to see where researchers may find potential collaboration opportunities, and we observe which authors have specialized or broad research interests. Research can be specialized but still integrate methods, techniques and data from multiple disciplines. We believe that an author who is represented primarily in a single cluster may not be engaging frequently in interdisciplinary research, or may be focusing on narrow research problems, or using similar research methods or equipment. In Figure 7 we see that the two astrochemists (Bennett and Kaiser) are entirely represented by cluster 8, consistent with the results presented in Figure 6. We know that their research is heavily influenced by their experimental apparati, thus suggesting that the experimental methods and apparati significantly affect the description of a research track. Interestingly, Schörghofer's research is on various planetary bodies such as Mars and the Moon, which is also true of Taylor. Therefore, clustering the text of the aggregated abstracts sufficiently illuminates similarities in research tracks across disciplinary boundaries, in this case, between astronomy and geology.

In Figure 8, we observe that Huss, Jewitt, Krot and Meech's research is found in many clusters. This signifies that their research is likely to be very interdisciplinary. With regards to those authors

represented by a few clusters, we cannot conclude that their research is absolutely mono-disciplinary, as it may be very specialized, or utilize the same methods or apparati. However, we believe that those UHNAI authors with publications in multiple clusters are *more likely* to be engaged in interdisciplinary research. In Figure 9, we observe that of the senior (non-postdoctoral fellows) astronomers (Reipurth, Meech, Jewitt, Haghighipour, Owen, Schörghofer) half (Meech, Jewitt, and Owen) are fairly diverse in their research interests and the other half (Reipurth, Haghighipour, Schörghofer) are engaged in specialized or mono-disciplinary research.

It is clear from these results that the sIB method in combination with our aggregated abstracts can illuminate where research areas overlap. Furthermore, while clusters do not inherently relate any information about a researcher's discipline, it is clear that researchers from the same department often cluster together. Therefore, we expect that performing a similar analysis on the entire NASA Astrobiology Institute will show where collaborations between researchers can occur, and can assist NASA with outlining research priorities. These results can be serve as the framework for a geospatial visualization of common yet unconnected research tracks and potential collaborators, similar to the "hot regions" described by Bornmann and Waltman (2011).

5 Discussion and Conclusions

We clustered astrobiology abstract data to evaluate SCs as document labels. We attempt to reconcile clustering (bottom-up approach) with pre-defined categories (top-down approach). The clusters produced by text mining the abstract data did not generally correspond well to the SCs. Therefore, we conclude that SCs are not well suited to the classification of astrobiology publications, and speculate that this may also be true for other interdisciplinary fields. Astrobiology research outputs cite mono-disciplinary and interdisciplinary publications which may prevent SCs from forming cohesive clusters. Additionally, as discussed in Small (2010), many journals publish highly diverse content, which no journal-level classification system could represent completely. The class imbalance problem in our dataset requires us to explore utilizing an oversampling technique. While we believe that the method remedies the skewed distribution of conflated SCs in our dataset, performing a text mining clustering analysis on a balanced astrobiology dataset without oversampling may produce different results. That is, SCs may be more accurate when the distribution of SCs is uniform with-

out oversampling using synthetic data. Nonetheless, the distribution of departmental affiliations of the UHNAI researchers is skewed, which affects the distribution of publications across different SCs; it is likely that this scenario will be consistent with the other NASA Astrobiology Institute teams.

Our results suggest that 10 clusters may be the most appropriate level at which to analyze the astrobiology collection (Figure 3). Too few clusters and the interdisciplinary diversity of the source documents is not well represented; too many and they may be oversegregated, lessening the chance to identify potential commonalities in documents from different disciplines and SCs. We suggest that when documents from different SCs cluster together, this may indicate implicit interdisciplinary connection, where knowledge in one field might inform another. These documents may warrant selective investigation by experts in each field, and suggest potentially productive interdisciplinary collaborations.

Similarly, text mining the aggregated abstracts using the sIB method is also suited to the task of finding collaboration opportunities. Our experiments consistently showed that authors from the same academic department tended to have their publications cluster together. If this were not the case, we would be unable to make any claims regarding the similarity of publications within a given cluster. Authors who are found in the same cluster denote potential collaboration opportunities. An author that has publications in many clusters indicates that they are engaged in interdisciplinary research, or perhaps that they are not, but should be. Those authors with few publications were either underrepresented in WoK and ADS, or were post-doctoral fellows at the UHNAI. We find that the UHNAI post-doctoral fellows are mostly engaged in IDR. This is an encouraging result, as promoting IDR is one of the goals of the NASA Astrobiology Institute. Younger generations of researchers will need to synthesize techniques from multiple disciplines to answer some of the most fundamental questions in science in general, and astrobiology in particular.

We insinuated that a strong conclusion cannot be made regarding those authors that are strongly represented in a single cluster. Research in this context is either: 1) interdisciplinary but specialized, perhaps incorporating a synthesis between methods, techniques and data from multiple disciplines, but with a narrow scope or 2) mono-disciplinary. Distinguishing between these two cases requires studying the individual words in each cluster. Additionally, such an analysis would lead to narrowing the scope of collaboration between two or more researchers that are found within a single

cluster. This analysis will be explored in future work.

The context of the interdisciplinary field of astrobiology has permitted us to explore a method of measuring interdisciplinarity, and identify potential collaboration opportunities. We find that most of the UHNAI team is engaged in IDR, and that our method suggests where interdisciplinary collaborations should occur. We believe our method, which combines bibliometrics and machine learning, makes valid predictions, based on our apriori knowledge of the structure of the research team and those intra-team collaborations that exist. Bibliometric studies of interdisciplinarity can benefit when augmented with machine learning algorithms, in an attempt to understand the fine-grained details of interdisciplinary research.

Acknowledgements

This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA08DA77A issued through the Office of Space Science.

References

Bjurström, A., & Polk, M. (2011). Climate change and interdisciplinarity: a co-citation analysis of IPCC Third Assessment Report. *Scientometrics*, 87, 525–550.

Bornmann, L., & Waltman, L. (2011). The detection of "hot regions" in the geography of science–A visualization approach by using density maps. *Journal of Informetrics*, 5(4), 547–553.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64, 351–374.

Brewer, G. D. (1999). The challenges of interdisciplinarity. *Policy Sciences*, 32, 327–337.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Cockell, C. (2002). Astrobiology–a new opportunity for interdisciplinary thinking. *Space Policy*, 18(4), 263–266.

Derrick, G., Sturk, H., Haynes, A., Chapman, S., & Hall, W. (2010). A cautionary bibliometric tale of two cities. *Scientometrics*, 84, 317–320.

Gargaud, M., & Tirard, S. (2011). Exobiology: An Example of Interdisciplinarity at Work. In Jean-Pierre Lasota, editor, *Astronomy at the Frontiers of Science*, volume 1 of *Integrated Science & Technology Program*, pages 337–350. Springer Netherlands.

Jacsó, P. (2005). As we may search: comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89, 1537–1547.

Kostoff, R. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, 43, 27–43.

Kostoff, R., del Río, J. A., Humenik, J. A., García, E. O., & Ramírez, A. M. (2001). Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, 52(13), 1148–1156.

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, 74, 273–294.

Morillo, F., Bordons, M., & Gómez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics*, 51, 203–222.

National Academies. Committee on Facilitating Interdisciplinary Research, of the Committee on Science, Engineering, and Public Policy. (2005). *Facilitating Interdisciplinary Research*. Washington, DC: National Academies Press.

National Science Foundation. Introduction to interdisciplinary research. `http://www.nsf.gov/od/oia/additional_resources/interdisciplinary_research/`, Accessed November 21, 2011.

Oliver, C. A., & Fergusson, J. (2007). Astrobiology: A pathway to adult science literacy? *Acta Astronautica*, 61(7-8), 716–723.

Porter, A., Cohen, A., Roessner, J. D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, 72, 117–147.

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81, 719–745.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82, 263–287.

Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136, New York, NY, USA.

Small, H. (2010). Maps of science as interdisciplinary discourse: co-citation contexts and the role of analogy. *Scientometrics*, 83, 835–849.

Staley, J. (2003). Astrobiology, the transcendent science: the promise of astrobiology as an integrative approach for science and engineering education and research. *Current Opinion in Biotechnology*, 14(3), 347–354.

Sugimoto, C. (2011). Looking across communicative genres: a call for inclusive indicators of interdisciplinarity. *Scientometrics*, 86, 449–461.

Upham, S., & Small, H. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, 83, 15–38.

van Raan, A. F. J., & van Leeuwen, T. N. (2002). Assessment of the scientific basis of interdisciplinary, applied research: Application of bibliometric methods in Nutrition and Food Research. *Research Policy*, 31(4), 611–632.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14–26.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques, 2nd edition*. Morgan Kaufmann, San Francisco.

Zhang, J., Vogeley, M. S., & Chen, C. (2011). Scientometrics of big science: a case study of research in the Sloan Digital Sky Survey. *Scientometrics*, 86, 1–14.

Zhang, L., Liu, X., Janssens, F., Liang, L., & Glänzel, W. (2010). Subject clustering analysis based on ISI category classification. *Journal of Informetrics*, 4(2), 185–193.

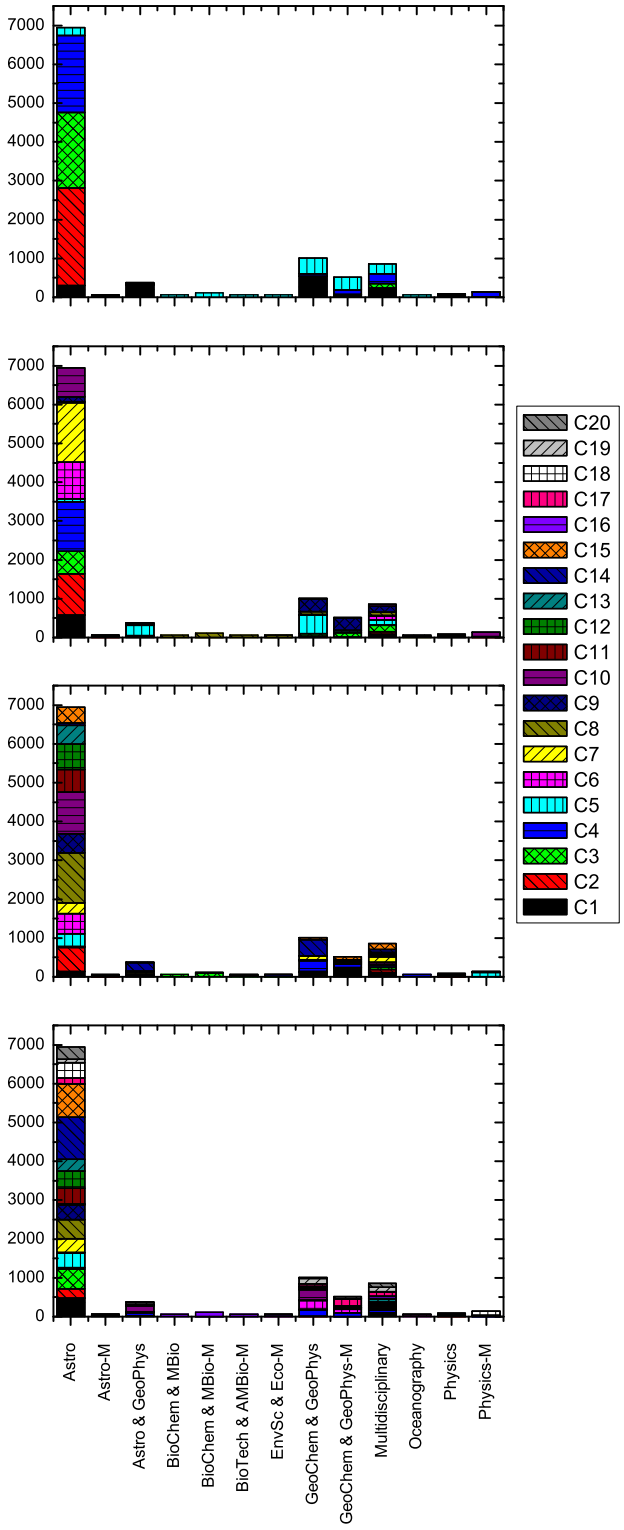

Figure 1: The results of clustering abstract data to evaluate subject categories as document labels. Results are given for 5, 10, 15 and 20 clusters from top to bottom.

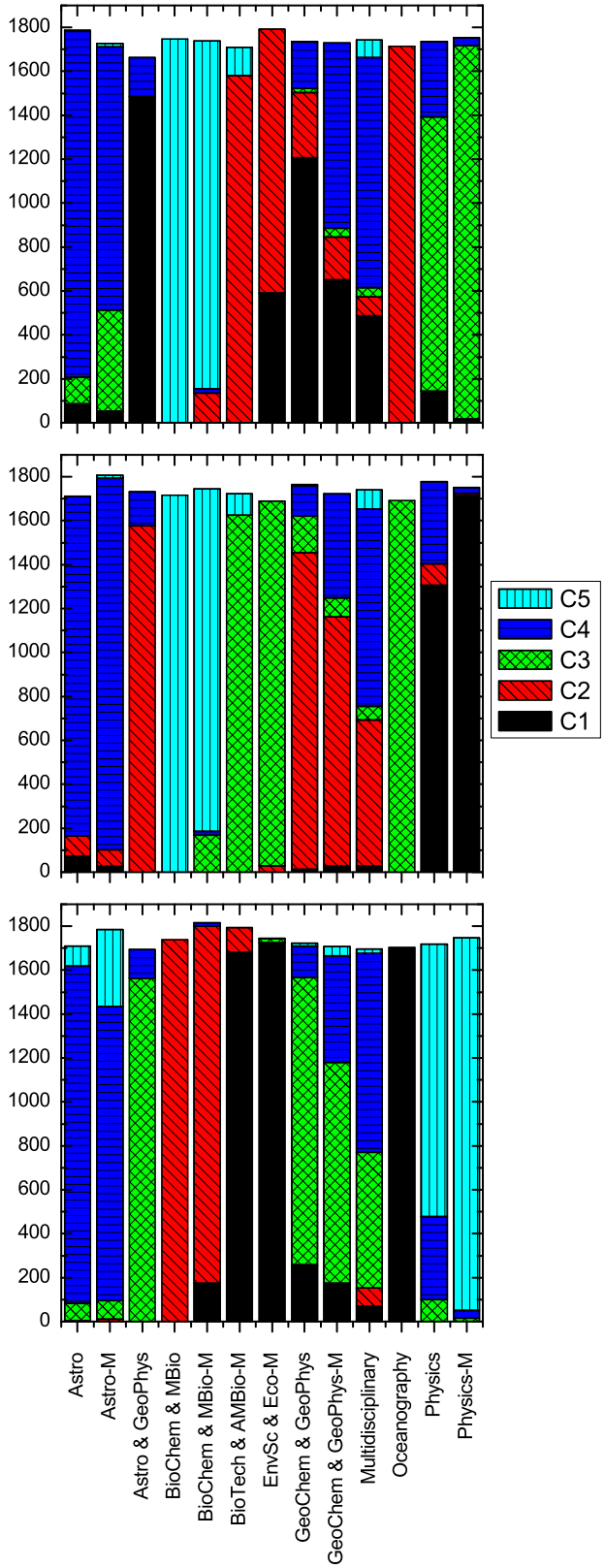

Figure 2: The results of clustering oversampled data in three separate trials, each representing 25% of the dataset. Each abstract is assigned one of five clusters.

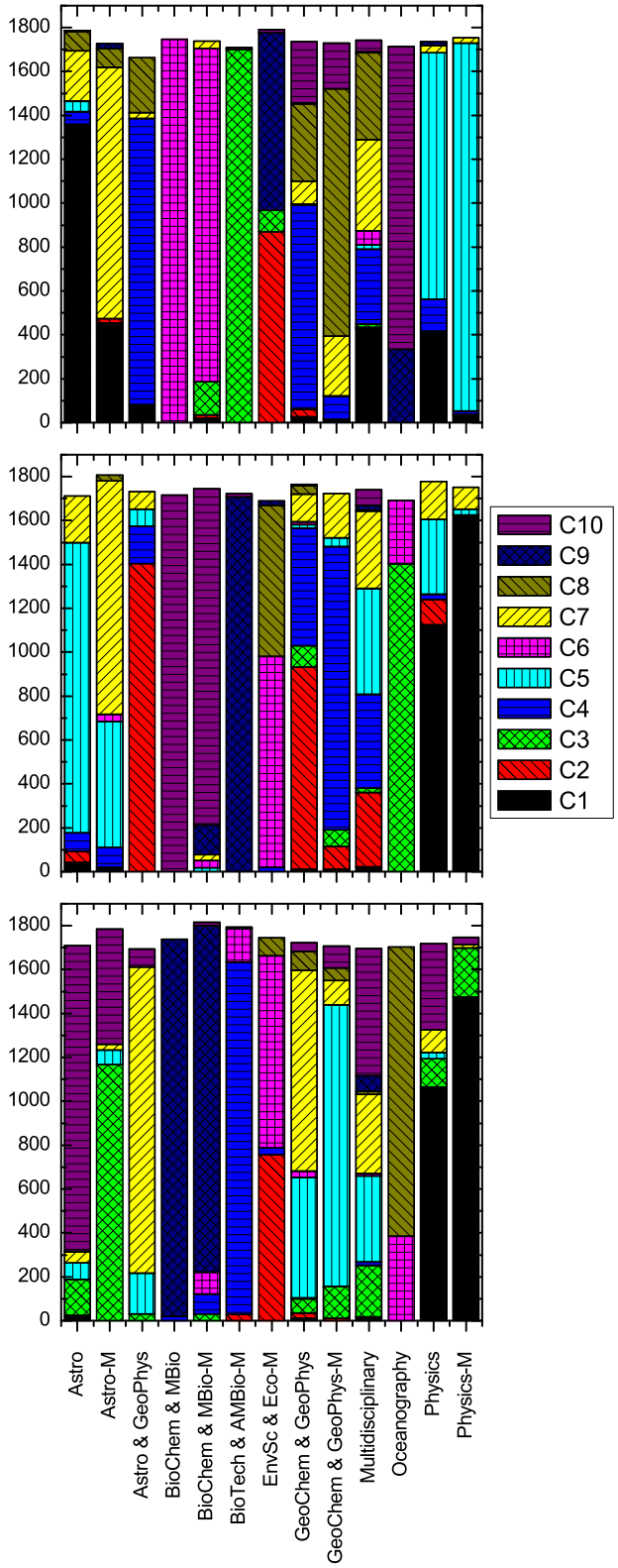

Figure 3: The results of clustering oversampled data in three separate trials, each representing 25% of the dataset. Each abstract is assigned one of ten clusters.

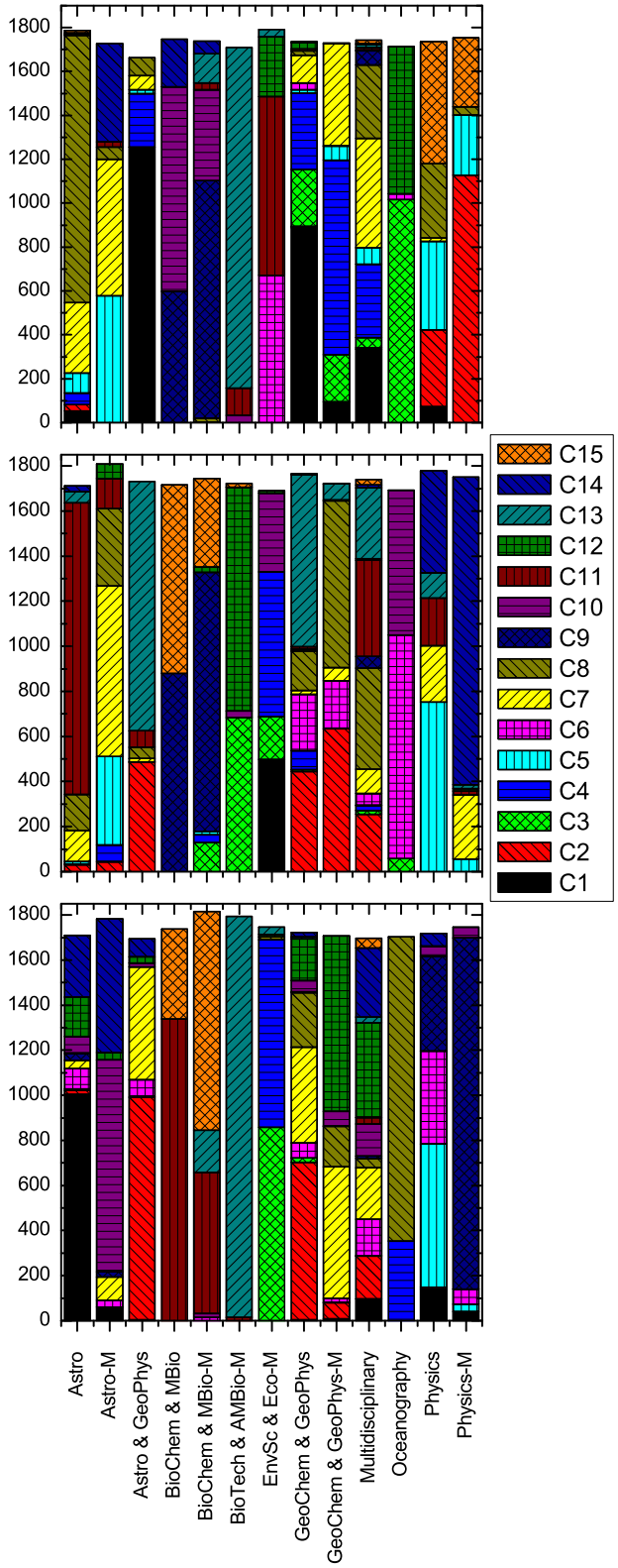

Figure 4: The results of clustering oversampled data in three separate trials, each representing 25% of the dataset. Each abstract is assigned one of fifteen clusters.

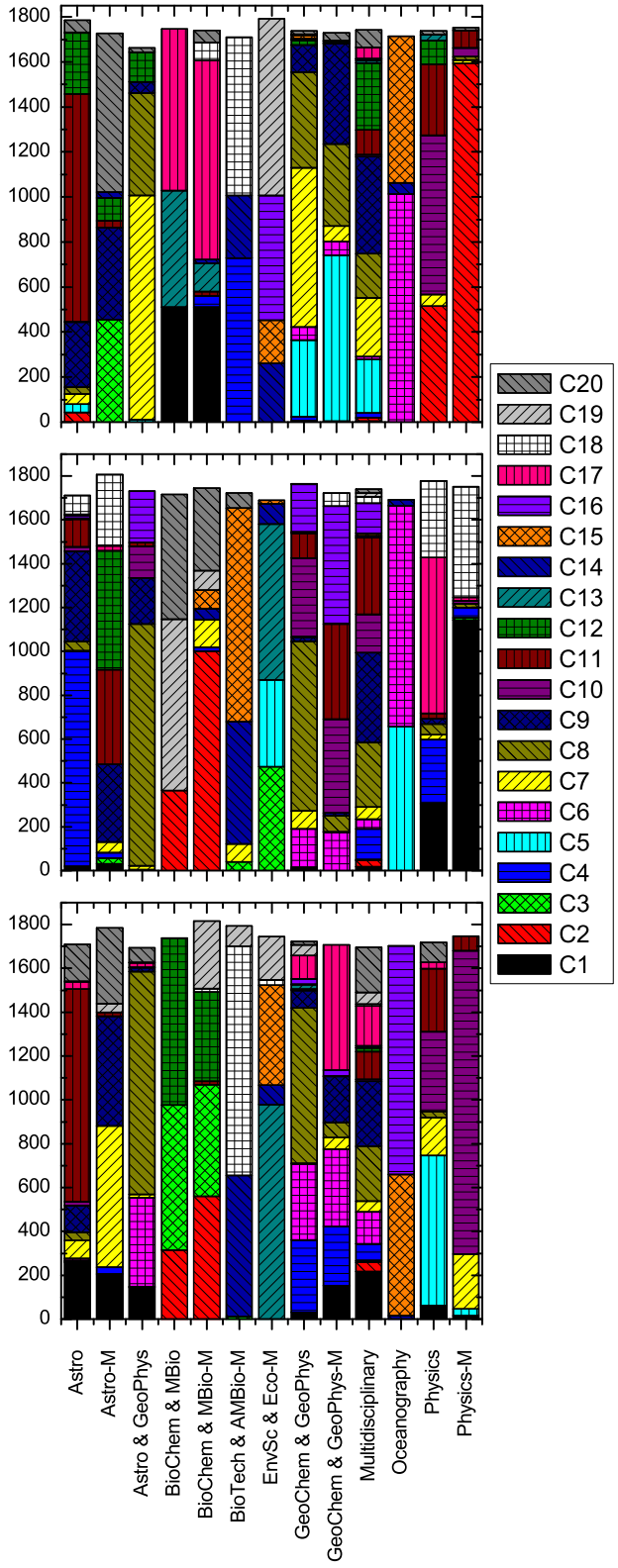

Figure 5: The results of clustering oversampled data in three separate trials, each representing 25% of the dataset. Each abstract is assigned one of twenty clusters.

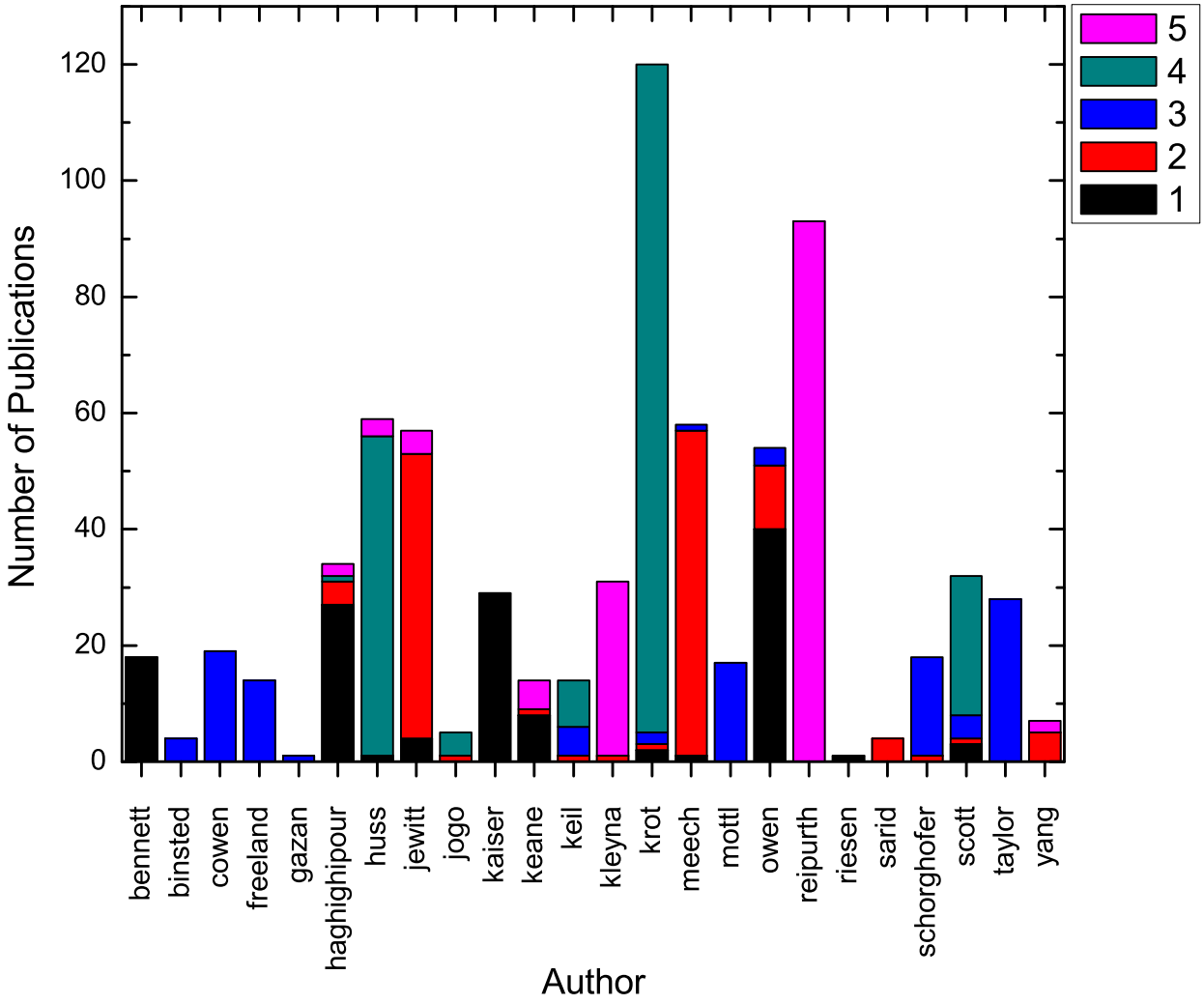

Figure 6: Clustering the aggregated abstracts using 5 clusters. This plot ensures that we are not obtaining extremely unlikely correlations and shows that researchers from the same academic department are largely clustering together. For example, Bennett is a post-doctoral fellow working with Kaiser; they often publish together and their aggregated abstracts are clustering entirely in cluster 1. As another example, the geologists/geophysists Krot, Keil, Huss, Scott and Jogo are all strongly represented in cluster 4. The one exception is Taylor, who appears to be clustering more strongly with the two oceanographers (Cowen and Mottl). As expected, researchers have the most in common with those in their home discipline.

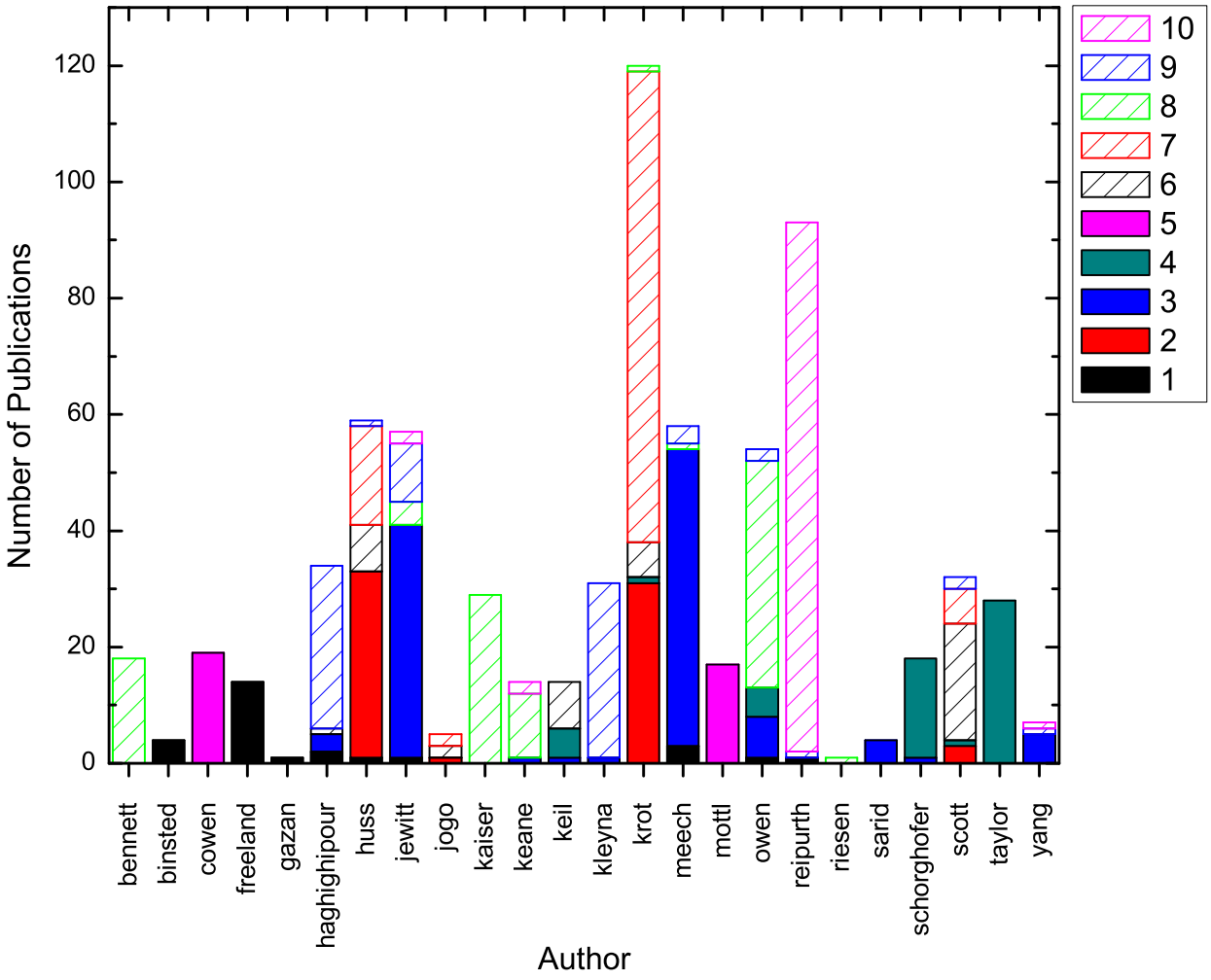

Figure 7: Clustering the aggregated abstracts using 10 clusters. The two oceanographers (Cowen and Mottl) have all of their papers clustering together in Cluster 5. The same is true of Bennett and Kaiser (Astrochemistry). In the previous figure, Taylor was clustering with the oceanographers. However, we can see here that Taylor's work is similar to that of Schörghofer's, despite their membership to different home disciplines (Geology, and Astronomy respectively). Rather striking is the mono-disciplinarity regarding Reipurth's research.

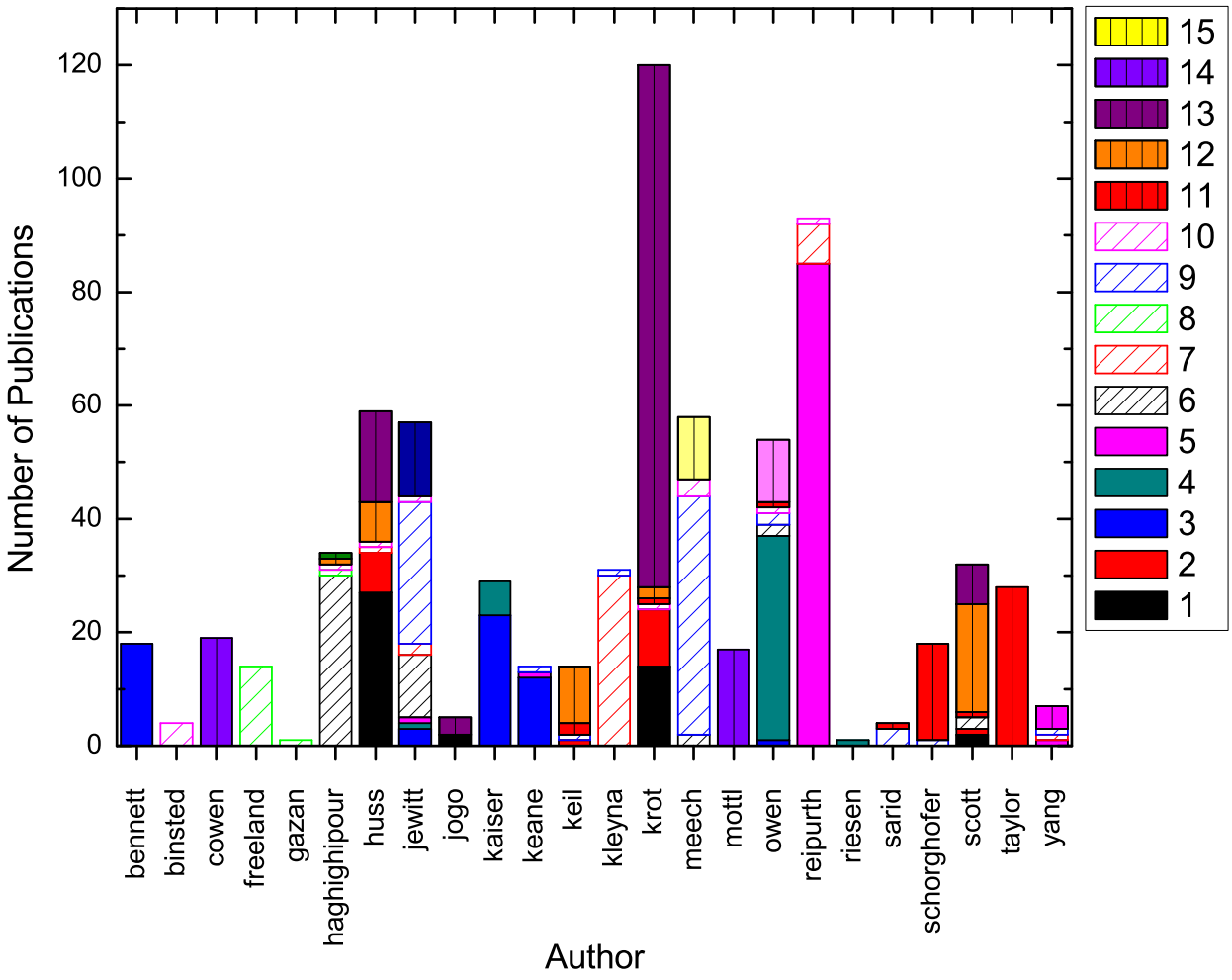

Figure 8: Clustering the aggregated abstracts using 15 clusters. In this figure, Bennett and Kaiser are no longer entirely represented by a single cluster. When we utilized 5 and 10 clusters, Binsted and Gazan (Computer Science) and Freeland (Biology) had their publications cluster together. We know in particular that the research by the computer scientists is likely to be the most dissimilar to all authors from other home disciplines. However, when clustering with 15 clusters, we observe Binsted's research depart from the cluster that contains Gazan and Freeland's research and that the research has a tangential relation to research produced by other team members.

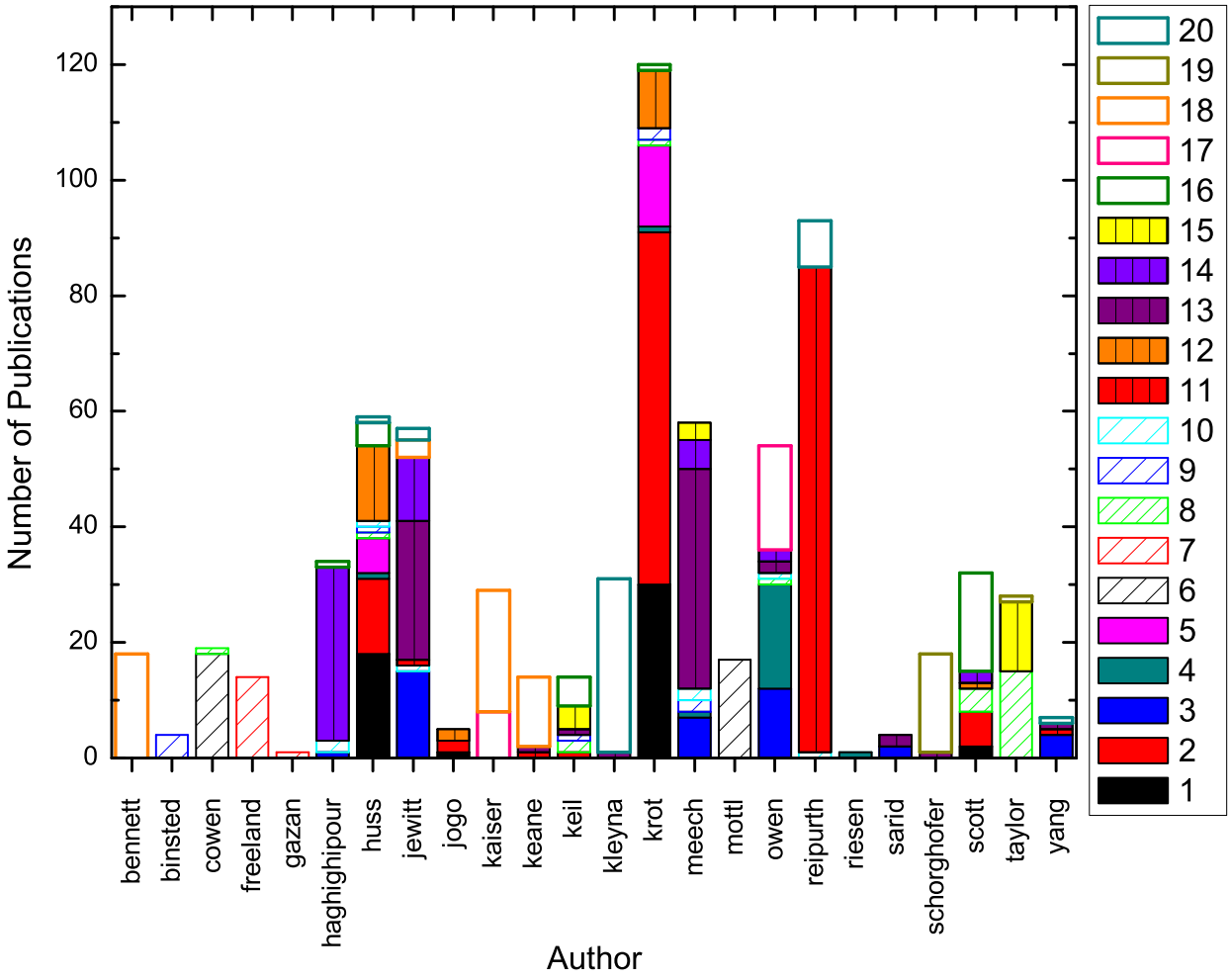

Figure 9: Clustering the aggregated abstracts using 20 clusters. If we assume that membership to many clusters indicates a high degree of interdisciplinarity, Huss is the most interdisciplinary UHNAI team member. Of the senior Astronomers (Reipurth, Meech, Jewitt, Haghighipour, Owen, Schörghofer) half (Meech, Jewitt, and Owen) are fairly diverse in their research interests, or engage in IDR, and the other half (Reipurth, Haghighipour, Schörghofer) are engaged in specialized or mono-disciplinary research. As younger researchers, the UHNAI post-doctoral fellows appear to be engaging in interdisciplinary research.

Table 1: The distribution of conflated subject categories and their corresponding abstracts.

Subject Category	Number of abstracts
Astronomy & Astrophysics [Astro]	6946
Astronomy & Astrophysics-Multidisciplinary [Astro-M]	66
Astrophysics & Geophysics [Astro & GeoPhys]	376
Biochemistry & Molecular Biology [BioChem & MBio]	62
Biochemistry & Molecular Biology-Multidisciplinary [BioChem & MBio-M]	109
Biotechnology & Applied Microbiology-Multidisciplinary [BioTech & AMBio-M]	58
Environmental Sciences & Ecology-Multidisciplinary [EnvSc & Eco-M]	68
Geochemistry & Geophysics [GeoChem & GeoPhys]	1009
Geochemistry & Geophysics-Multidisciplinary [GeoChem & GeoPhys-M]	522
Multidisciplinary Sciences [Multidisciplinary]	853
Oceanography	66
Physics	86
Physics-Multidisciplinary [Physics-M]	138

Table 2: Home discipline of the authors at the University of Hawaii NASA Astrobiology Institute. An asterisk (*) denotes a post-doctoral researcher.

Author	Departmental Affiliation/Home Discipline
Bennett*	Chemistry
Binsted	Computer Science
Cowen	Oceanography
Freeland	Biology
Gazan	Computer Science
Haghighipour	Astronomy
Huss	Geology
Jewitt	Astronomy
Jogo*	Geology
Kaiser	Chemistry
Keane*	Astronomy
Keil	Geology
Kleyna*	Astronomy
Krot	Geology
Meech	Astronomy
Mottl	Oceanography
Owen	Astronomy
Reipurth	Astronomy
Riesen*	Astronomy
Sarid*	Astronomy
Schörghofer	Astronomy
Scott	Geology
Taylor	Geology
Yang*	Astronomy